Probing Linguistic Features of Sentence-Level Representations in Neural Relation Extraction

ACL 2020

Christoph Alt, Aleksandra Gabryszak, Leonhard Hennig

German Research Center for AI (DFKI)
Speech and Language Technology Lab

Relation Extraction

Relation extraction (RE) is concerned with extracting semantic relations from text

?

The measures included $\underbrace{\text{Aerolineas}}_{\text{head}}$'s domestic subsidiary, $\underbrace{\text{Austral}}_{\text{tail}}$.

org:subsidiaries

org:parents

org:members

. . .

Neural network-based models have considerably improved RE performance

[Baldini Soares et al., 2019; Peters et al., 2019; Joshi et al., 2019; Li et al., 2020]

But, what do neural network-based models consider relevant for relation prediction?

Motivation

- **Goal:** Understand what aspects of the input neural RE models consider relevant for a prediction
  - Gain further insights into decision process
    - → Identify areas for improvement
  - Crucial to ensure accountability, trust, and fairness
    - → important in critical domains, e.g., healthcare
- **Problem:**
  - Nested non-linear structure makes neural networks highly non-transparent
  - Un- or self-supervised pre-training made models even more complex

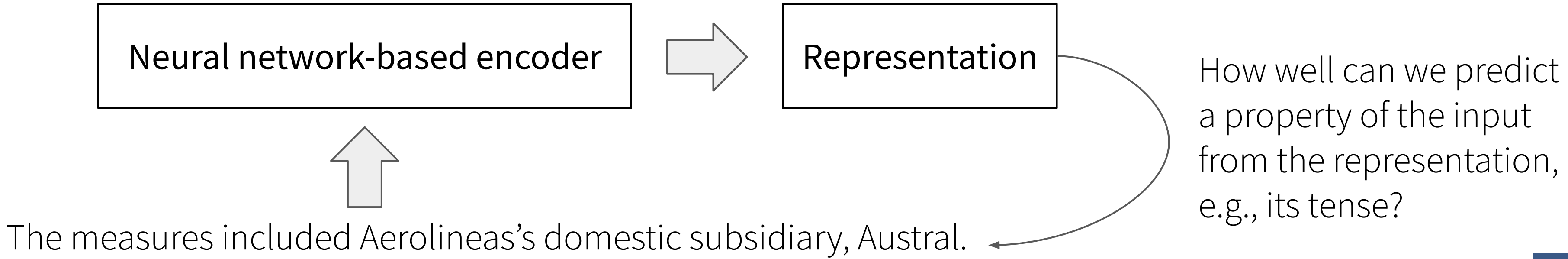
Research Questions

- What linguistic properties are encoded by neural RE models?
  - How well do models encode well known features for RE?
  - How does neural network architecture affect the captured features?
  - How does additional linguistic information affect the encoded features?
  - How does this affect performance on the RE task?

Sentence Level Probing Tasks

- Probing task [Adi et al., 2017], diagnostic classifier, or auxiliary prediction task
  - Classifier trained on a set of model's internal representations
  - Performance measures how well the information is encoded
    - → Assumption: Information is used for model prediction

Example: Probing of a general sentence encoder [Conneau et al., 2018]



Linguistic Probing Tasks for Neural RE

- Set of probing tasks for RE → Features that proved useful in earlier work
- Surface, syntactic, and semantic properties of sentences with marked entities
  - Sentences collected from TACRED [Zhang et al., 2017] and SemEval 2010 Task 8 [Hendrickx et al., 2010]

Category	Properties
Surface	■ Sentence length ■ Argument distance → number of tokens between mentions ■ Named entity exists between mentions
Syntactic	■ Dependency tree depth ■ Shortest dependency path (between mentions) tree depth ■ Argument order → whether head comes before tail ■ Part of speech of tokens to the left and right of {head, tail}
Semantic	■ Named entity type of {head, tail} ■ Grammatical role of {head, tail}

Experimental Setup

- Datasets: TACRED and SemEval 2010 Task 8
- Evaluate probing tasks on trained RE models of different architectures
  - Baseline: Bag of embeddings
  - CNN
  - Bi-LSTM
  - GCN (Graph convolution)
  - Self-attention
- Combined with supporting linguistic knowledge
  - Entity masking
    - → i.e., replacing entity mentions with named entity type
  - Contextual word representations
    - BERT
    - ELMo

General Probing Task Performance

	Type Head	Type Tail	Sent Len	Arg Dist	Arg Ord	Ent Exist	PosL Head	PosR Head	PosL Tail	PosR Tail	Tree Dep	SDP Dep	GR Head	GR Tail	F1 score
Majority vote	66.4	33.5	14.5	14.8	54.7	51.0	22.8	23.0	26.9	20.0	23.7	28.4	58.4	75.2	-
Length	66.4	33.5	**100.0**	13.8	54.8	59.4	18.6	24.7	26.9	20.1	**30.5**	29.6	58.4	75.2	-
ArgDist	66.4	33.5	16.5	**100.0**	54.7	77.5	14.9	23.0	26.9	19.8	23.8	35.3	58.4	75.2	-
BoE	77.7	47.6	61.1	22.6	97.3	66.5	33.7	41.5	32.5	36.3	29.8	31.0	66.3	77.4	39.4
CNN ⊗	84.2	60.9	46.4	58.3	94.3	81.5	44.3	50.9	54.4	63.9	27.7	40.0	68.5	82.0	59.5
+ BERT ↑	87.2	79.3	50.6	25.3	78.3	69.8	39.6	42.9	59.9	77.5	30.3	35.1	65.6	86.9	66.1
GCN ⊗	87.6	67.4	18.1	33.1	81.6	72.8	36.8	51.1	44.8	48.8	24.1	47.3	73.2	83.0	63.7
+ BERT ↑	93.4	72.0	23.7	33.2	90.4	73.9	42.8	50.1	44.0	48.3	24.9	48.0	72.9	83.0	65.9
S-Att. ⊗	79.5	56.5	29.0	44.3	91.2	79.5	29.6	43.0	36.1	60.3	26.1	39.6	64.7	79.5	65.9
+ BERT ↑	80.0	69.0	31.9	32.8	78.6	76.6	30.3	34.2	37.5	39.2	27.0	38.2	60.4	79.9	**66.9**

- Compared to baselines
  - all encoders perform superior on entity type tasks
  - all encoders perform lower on sentence length task
- GCN performs best on SDP tree depth

Effect of Neural Network Encoder Architecture

	Type Head	Type Tail	Sent Len	Arg Dist	Arg Ord	Ent Exist	PosL Head	PosR Head	PosL Tail	PosR Tail	Tree Dep	SDP Dep	GR Head	GR Tail	F1 score
CNN ⊗	84.2	60.9	46.4	58.3	94.3	81.5	44.3	50.9	54.4	63.9	27.7	40.0	68.5	82.0	59.5
Bi-LSTM ⊗	81.9	71.4	27.6	35.6	90.6	73.2	36.1	40.5	59.3	66.4	25.7	38.4	64.6	85.3	62.9
GCN ⊗	87.6	67.4	18.1	33.1	81.6	72.8	36.8	51.1	44.8	48.8	24.1	47.3	73.2	83.0	63.7
S-Att. ⊗	79.5	56.5	29.0	44.3	91.2	79.5	29.6	43.0	36.1	60.3	26.1	39.6	64.7	79.5	65.9
CNN	94.0	85.8	47.6	88.1	98.8	**84.5**	70.7	76.1	84.0	86.5	28.5	44.0	78.0	88.6	55.9
Bi-LSTM	93.4	81.2	42.0	47.9	**99.4**	79.2	41.2	50.8	50.6	68.4	28.7	41.7	69.3	85.2	55.3
GCN	93.0	81.9	18.8	35.5	86.0	74.4	48.6	48.8	51.2	52.3	24.0	**49.9**	74.2	85.9	57.4
S-Att.	89.9	81.8	22.7	32.8	75.7	78.1	34.1	38.9	40.8	44.8	26.1	38.2	60.7	81.1	57.6

- Models with a local or recency bias, e.g., CNN, Bi-LSTM
  - perform well on probing tasks with local focus
  - perform well on distance related tasks
- Models with access to dependency information (GCN)
  - perform well on tree related tasks
- Self-attention superior RE performance but consistently lower on the probing tasks

Effect of Contextual Word Representations

	Type Head	Type Tail	Sent Len	Arg Dist	Arg Ord	Ent Exist	PosL Head	PosR Head	PosL Tail	PosR Tail	Tree Dep	SDP Dep	GR Head	GR Tail	F1 score
CNN	94.0	85.8	47.6	88.1	98.8	**84.5**	70.7	76.1	84.0	86.5	28.5	44.0	78.0	88.6	55.9
+ BERT ↑	96.1	88.8	48.0	43.7	91.9	80.0	56.9	70.3	80.1	87.5	28.0	41.3	75.0	89.6	61.0
CNN ⊗	84.2	60.9	46.4	58.3	94.3	81.5	44.3	50.9	54.4	63.9	27.7	40.0	68.5	82.0	59.5
+ BERT ↑	87.2	79.3	50.6	25.3	78.3	69.8	39.6	42.9	59.9	77.5	30.3	35.1	65.6	86.9	66.1
S-Att.	89.9	81.8	22.7	32.8	75.7	78.1	34.1	38.9	40.8	44.8	26.1	38.2	60.7	81.1	57.6
+ BERT ↑	96.5	87.3	26.1	32.6	76.8	78.0	34.7	40.9	40.0	44.0	25.7	38.1	62.2	81.7	63.8
S-Att. ⊗	79.5	56.5	29.0	44.3	91.2	79.5	29.6	43.0	36.1	60.3	26.1	39.6	64.7	79.5	65.9
+ BERT ↑	80.0	69.0	31.9	32.8	78.6	76.6	30.3	34.2	37.5	39.2	27.0	38.2	60.4	79.9	**66.9**
+ BERT ↓	82.4	66.9	36.2	33.2	74.9	76.8	32.0	37.6	38.0	41.3	27.4	37.6	63.0	79.8	66.7

- Contextual word representations increases performance on entity type and POS related tasks
- Uncased BERT performs equal or better on named entity and POS tasks
- Leads to overall increase in RE performance

Conclusion

- Extensive evaluation showed that
  - self-attentive encoders are well suited for RE
  - but perform lower on probing tasks
  - bias induced by different architectures is reflected in probing task performance
    - e.g., distance and dependency related tasks
- However, probing task performance *not correlated with RE performance*
- **Software libraries:**
  - REval, framework extending SentEval [Conneau and Kiela, 2018] to develop and eval. RE probing tasks
  - RelEx, binary RE framework based on AllenNLP [Gardner et al., 2017]

Thank you

Github: https://github.com/DFKI-NLP/REval

References

- Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. Position-aware attention and supervised data improve slot filling. EMNLP, 2017.
- Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid O Seaghdha, Sebastian Pado, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. SemEval, 2010.
- Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. ACL, 2019.
- Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. Knowledge enhanced contextual word representations. EMNLP, 2019.
- Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, Hannaneh Hajishirzi. A General Framework for Information Extraction using Dynamic Span Graphs. NAACL, 2019.
- Cheng Li, Ye Tian. Downstream Model Design of Pre-trained Language Model for Relation Extraction Task. arxiv:2004.03786, 2020.
- Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke S. Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. TACL, 2019.
- Alexis Conneau, German Kruszewski, Guillaume Lample, Loic Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. ACL, 2018.
- Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. ICLR, 2017.
- Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. LREC, 2018.
- Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew E. Peters, Michael Schmitz, and Luke Zettlemoyer. Allennlp: A deep semantic natural language processing platform. arXiv:1803.07640, 2017.